DOCUMENT RESUME

ED 407 423 TM 026 445

AUTHOR Kier, Frederick J.
TITLE Ways To Explore the Replicability of Multivariate Results (Since Statistical Significance Testing Does Not).
PUB DATE 23 Jan 97
NOTE 17p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 23-25, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Methods; *Multivariate Analysis; *Sampling; *Statistical Significance
IDENTIFIERS Bootstrap Methods; Cross Validation; Jackknifing Technique; *Research Replication

ABSTRACT

It is a false, but common, belief that statistical significance testing evaluates result replicability. In truth, statistical significance testing reveals nothing about results replicability. Since science is based on replication of results, methods that assess replicability are important. This is particularly true when multivariate methods, which capitalize on sampling error, are used. This paper explores three methods that can give an idea of the replicability of results in multivariate analysis without having to repeat the study. The first method is cross validation, a replication technique in which the entire sample is first run through the planned analysis and then the sample is randomly split into two unequal parts so that separate analyses are done on each half. The jackknife is a second method of replicability that relies on partitioning out the impact or effect of a particular subset of the data on an estimate derived from the total sample. The bootstrap, a third method of studying replicability, involves copying the data set into an infinitely large "mega" data set. Many different samples are then drawn from the file and results are computed separately for each sample and then averaged. The main drawback of all these internal replicability procedures is that their results are all based on the data from the one sample being analyzed. However, internal replication techniques are better than not addressing the issue at all. (Contains 18 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 407 423

Ways to Explore the Replicability of Multivariate Results

(Since Statistical Significance Testing Does Not)

Frederick J. Kier

Texas A&M University 77843-4225

Running head: MULTIVARIATE REPLICABILITY

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
FREDERICK J. KIER
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 23, 1997.

TM026445

Abstract

It is a false, yet somewhat common, belief that statistical significance testing evaluates result replicability. Since, in truth, statistical significance testing reveals absolutely nothing about result replicability, and since science is based upon replication of results, methods that do assess replicability are important. This is particularly true when using multivariate methods, which capitalize on sampling error. This paper explores three methods (cross-validation, jackknife, and bootstrap) that can be used to get an idea of the replicability of one's results in multivariate analyses, without actually having to perform a study again.

Ways to Explore the Replicability of Multivariate Results

(Since Statistical Significance Testing Does Not)

As discussed by many recent authors (Fish, 1988; Thompson, 1994a; Stevens, 1996; Hinkle, Wiersma, & Jurs, 1994), multivariate methods are becoming nearly mandatory to use in the social sciences. Thompson (1994a) stated that there are two reasons why multivariate methods are usually vital. First "multivariate methods limit the inflation of Type I 'experimentwise' error" (Thompson, 1994a, p. 9). Experimentwise error increases when a researcher uses multiple univariate analyses (such a t-tests, ANOVAs, etc.) instead of using a multivariate analysis. Each individual univariate analysis adds to the chance that one of these analyses will be due to error, hence, the aforementioned inflation of Type I "experimentwise" error [for a more detailed discussion of this issue, the reader is urged to consult Thompson (1994a) or Fish (1988)]. Second, as Thompson (1994a) also states "multivariate methods best honor the reality to which the researcher is purportedly trying to generalize" (p. 12). In others words, since, in reality, variables are often (if not always) influenced by, or correlated with, many other variables, multivariate analyses are a better fit to the "real world" which we are investigating.

Although using multivariate analyses gives us the aforementioned advantages, like their univariate brethren, they still tell us nothing about the replicability of our results. It is a common, but very false, myth that statistical significance testing gives an indication of the likelihood of replication (Cohen, 1994; Thompson, 1996). Statistical analyses per se tell us nothing about replicability either. Since all such analyses are correlational, they tell us the relationships between variables, but nothing about replicability (Knapp, 1978). The

bottom line is that only way to get an idea about the likelihood of one's results replicating, without drawing another sample and actually re-doing the study, is to perform a replicability analysis of some kind.

Why is replication important? As Thompson (1996) stated: "If science is the business of discovering replicable effects, because statistical significance tests do not evaluate result replicability, then researchers should use and report some strategies that *do* evaluate the replicability of their results" (p. 29). Thompson (1996) also stated that actually re-performing the study with a new sample ("external replication") is the only way to directly assess replicability. There are, however, several methods a researcher can use that do not involve the sometimes heavy work needed to re-perform a study. These are frequently referred to as "internal replication" methods, and the use of three of these methods, cross-validation, jackknife, and bootstrap, with multivariate analyses will be the focus of the present paper.

Apart from this philosophy-of-science rationale for replication, there are statistical considerations that warrant the need for replication as well. King (1997) noted that "Each sample collected from a population of interest will yield at least slightly different results from any other independent sample. Thus, two researchers can potentially draw similiar samples and yet infer diverse theories based on their data" (p. 2). The only way to avoid this is to re-perform the study and get results from several samplings of the population (i.e., "external" replication). The so-called internal replication techniques described in this paper do not eliminate this error, but do give the researcher at least some idea of the replicability of the results. King (1997) made the important point that sampling error can

not be eliminated via random sampling procedures, due to the sample-specificity of the statistics calculated from the sample.

Replication is of particular importance in multivariate analyses (particularly canonical correlation analysis) because these analyses offer even more opportunities to capitalize on sampling error. In other words, they give a "worst case scenario" in terms of the effects of sampling error on determining the differences between the groups (or whatever we are trying to analyze) in question. Thus, replication analyses, at the least internal, if not external, replication analyses, are important, if not critical, in studies using multivariate analyses.

As mentioned earlier, the present paper focuses only on the three most common "internal" replication methods: cross-validation, jackknife, and bootstrap. These are not the only internal replication techniques, but these three are the most common and arguably the easiest to implement and use. We will describe each of these methods in a general, step-by-step process, using examples from the research literature as guides, with emphasis on the process of performing the technique. This is so the reader interested in using a particular method with another type of multivariate analysis can still follow the general guidelines. The following discussion also assumes that the reader has some knowledge of multivariate analyses, as a full discussion of this topic is beyond the scope of the present paper.

Cross-Validation

Crossman (1996) discusses how to perform a cross-validation on a canonical correlation analysis (CCA). CCA looks at the correlational relationships between two sets

of variables (dependent and independent, sometimes referred to as criterion and predictor, respectively), there must be at least two criterion and two predictor variables in each set, and these sets **must be meaningful** (Thompson, in press).

In the analysis, CCA first computes the correlations of the variables in the form of quadrants, each of which is associated with the correlations between variables in their variable sets. CCA then computes the quadruple product matrix, computed from these quadrants, and a principal components analysis is performed on this matrix (Thompson, in press). This results in standardized canonical function coefficients, which "are directly akin to beta weights in regression" and canonical structure coefficients (comparable to structure coefficients in regression) (Thompson, in press). These are related via functions (again akin to regression equations), the number of which equals the number of variables in the smaller variable set (Thompson, in press).

Cross-validation is a replication technique where one's entire sample is first run through the analysis (in our case a CCA) and then the sample is randomly split into two unequal parts and then separate analyses (CCA's) are done on each half. Unequal subsample sizes are used in order that the researcher will be able to discern the two groups (when the subsamples are of equal size, it is easy to confuse which group is which). The key to cross-validation in its use with CCA is that "new predictor and criterion composite scores for the first group are derived from standardized function coefficients of the second group" (Crossman, 1996, p. 8) and vice-versa. Note that the term "standardized" here means that the function coefficients are applied to measured variables in z-score form.

Cross-validation estimates the amount of "shrinkage" of the correlation

coefficients when the function coefficients of other subsample(s) are used (Crossman, 1996). The amount of "shrinkage" indicates the replicability of the results. It is important, however, to remember the adage "square before you compare" as the amount of variance is reflected by r^2 not r, and thus one must square these invariance statistics before one can compare them, or compute a meaningful difference.

It should be emphasized that replication techniques (like cross-validation) are especially (if not critically) important to use when performing a CCA. All multivariate analyses capitalize on sampling error, but CCA is particularly susceptible to biases in one's sample. It should also be mentioned that one needs a very large sample size in order to do a cross-validation with a CCA. The CCA itself requires a large sample size, and, with a cross-validation, one must do two more CCA's on each subsample. The main advantage of using a cross-validation over the jackknife and bootstrap techniques is that this method is relatively simple to implement, and can be implemented using many of the statistical computer packages on the market. Its drawback is that it is not as good of a measure of one's result replicability as the other two techniques.

Jackknife

The "jackknife" is another method of replicability that can be applied to multivariate analysis, such as a descriptive discriminant analysis (DDA). The jackknife technique was developed by Quenouille (1949) and Tukey (1958). Crask and Perreault (1977) stated "the essence of the jackknife approach is to partition out the impact or effect of a particular subset of the data (e.g., a single case) on an estimate derived from the total

sample" (p. 61). In other words, jackknife tries to control for a "piece" of your sample which may be exerting too much influence on your results due to sampling error. Although we will limit our discussion of jackknife to assessing the replicability of the results of multivariate analyses, specifically DDA, it should be noted that the jackknife can be used in many other domains and can be an extremely useful tool.

Daniel (1989) gives an excellent example of how to perform a jackknife on a DDA, and the basic procedure is similiar for using the jackknife to assess replicability in any other analysis. First, one performs the DDA as one normally would. Then you must divide up your sample into subsets. Usually these subsets have m (size of the subsets) equal to one, but one can pick any size for one's subsets, as long as the equation, n (total sample size) = k (number of subsets) * m (size of the subsets) (Daniel, 1989). Any predictive estimator, in the case of a DDA, a discriminant function coefficient, is then computed using all of the subsets (i.e. the entire sample). This predictive estimator, calculated using the entire sample, is referred to by Daniel (1989) as theta-prime. The same estimator (again, in the case of a DDA, a discriminant function coefficient) is again computed for the whole sample minus one of the subsets, and this is repeated for each subset, and this value is referred to by Daniel (1989) as theta. Psuedovalues are then computed by multiplying the number of subsets by theta-prime, and subtracting the number of subsets minus one, multiplied by theta. The jackknifed estimator is the average of these psuedovalues (Daniel, 1989).

After finding the jackknifed estimator, one performs a *t*-test on it (or one can compute a confidence interval for the jackknifed estimator). Divide the jackknifed estimator by its standard error to obtain a *t*-value, with degrees of freedom equal to k - 1. A jackknifed estimator is considered stable (i.e., your results are more likely to be replicable) if its calculated *t*-value exceeds the *t*-critical value (Daniel, 1989).

A conceptual illustration of what the jackknife does can be found in what the author calls "the sausage example." Say you are a making sausage and you have a big vat of sausage being made, from which represents the population you are sampling. Say that a bug accidentally gets mixed in with your sausage. The bug represents an extreme outlier that exists in your population. You take a sample from your sausage vat, a round, one-foot long sausage, which represents your sample taken from the population (the big vat) and lets say that, through sampling error, you get the bug in this particular sausage (i.e., this sample contains the outlier). What can we do about this "outlier" in our "sausage"? Well, using a knife (or even a jackknife!), we could cut our sausage into several pieces, in order to determine if one of the pieces has the bug, the "outlier" in it. In a similar fashion, the statistical jackknife lets us know when we have a problem with part of our sample due to sampling error.

Fan and Wang (1995) discuss some of the limitations of the jackknife appraoch to internal replication. Due to the fact that sample size does impose a limit on the number of resamples, the jackknife may not be appropriate for small samples. Fan and Wang (1995) also stated:

> it is still unclear whether, for a given sample, the size for each of the K subsets will cause any systematic differences in the results. In other words, does it matter if one observation is deleted for each jackknife analysis compared to five observations deleted each time? (p. 5)

Fan and Wang (1995) compared the jackknife to the bootstrap, and found that the two pretty much gave similar results when sample size was large. Thus, the jackknife does not appear to be the best method to use when sample size is small. Since cross-validation, especially in a CCA, also requires a large sample size, it is recommeded that the bootstrap, described below, be used in lieu of either the jackknife or cross-validation when the sample size is small.

Bootstrap

The bootstrap technique for determining result replicability was originally formulated by Efron (1979). Thompson (1995) described the conceptual basis of the technique:

> Conceptually, these methods involve copying the data set many times into an infinitely large "mega" data set. Then hundreds or thousands of different samples are drawn from the "mega" file and results are computed separately for each sample and then averaged. The method is powerful because the analysis considers so many configurations of subjects (including configurations in which a subject may be represented several times or not at all) and informs the researcher regarding the extent to which results generalize across different types of subjects. (p. 86)

Although conceptually rather simple, the bootstrap is a powerful technique, that, unfortunately, is difficult to perform on many conventional statistical computer packages. The step-by step conceptual instructions on how to perform the bootstrap will be given below, but it is recommended that the researcher who is seriously considering using this technique acquire a program [such as Thompson's CANSTRAP program, for applying bootstrap to canonical correlation analyses; see Thompson, (1995)].

We will be following the steps to use the bootstrap on a CCA, following Thompson's (1995) example. The first step is to perform the analysis as one normally would, in our case, a CCA. The second step involves the creation of a target matrix. This space creates a common function space so that the function is the same function in all our subsequent resamples. King (1997) states: "Only when functions remain constant across resampling can one legitimately compare results from the multiple samples" (p.13). The purpose of the target matrix is to make this so. One can create such a matrix using either structure or canonical function coefficients matrix from your sample at hand, or by creating one based on previous research or theory (Thompson, 1995). Whichever one of these you choose, however, must be used throughout the bootstrap procedure.

The next step involves resampling with replacement. It is important to remember that these resamples be the same size as your original sample. Thompson (1995) stated the reason for this is to "mimic the influences of the actual sample size" (p. 88). The last step is to perform a Procrustean rotation of each of the resamples. Again remember that you must rotate the same type of matrix that your target matrix space is.

Looking at the results of the "bootstrapped" CCA, one determines if the results are replicable by looking at the mean R_c^2 to the standard error of R_c^2. If this ratio is greater than 2, then your results are likely to be stable, or replicable. Like any analysis involving a distribution, however, it is also important to look at the values that describe the distribution, for example, the shape of the distribution. In a "bootstrapped" CCA, for example, one will find that the shape of the R_c^2 distribution over the 1,000 resamples will be positively biased. This is due to the fact that CCA capitalizes on sampling error (Thompson, 1995).

Conclusion

The main drawback of all these internal replicability procedures is that their results are all based on the data from the one sample being analyzed (King, 1997). King (1997) recommended using more than one of the internal replication techniques, and only when it is difficult or impossible to draw new samples and do a true, "external" replication. However, most researchers would likely balk at having to re-perform their studies, and internal replication techniques offer a way of at least getting some idea of the replicability of one's results. Internal replication techniques are better than not addressing the issue at all, which is presently a very common occurrence in the research literature. One reason may be that the myth that statistical significance testing indicates result replicability still permeates the thinking of many researchers. Another reason may be that many researchers do not do anything further after performing statistical significance tests because that is all they need to do to get their results published. The movement to limit, or even abolish, statistical significance testing, may aid in decreasing this attitude. A

movement to promote the use of internal replication techniques, especially in multivariate analyses, should also be undertaken. Replication techniques are more critical in multivariate analyses (particularly CCA) because these analyses capitalize on sampling error, giving a "best case scenario". Replication techniques would also likely be more frequently used if statistical computer packages featured these analyses.

Of course, this last recommendation comes with the problem of researchers performing "knee-jerk" analyses, without first thinking about what they are doing and why they are doing it. Internal replication analyses, like statistical significance testing, do not absolve the researcher from having to think about and make judgments about a study's results. They do, however, add an important, all too often overlooked, element to one's study.

References

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Crask, M. R., & Parreault, W. D. (1977). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-67.

Crossman, L. L. (1994, April). Cross-validation analysis for the canonical case. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 367 722)

Daniel, L. G. (1989, January). Use of the jackknife statistic to establish the external validity of discriminant analysis results. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 305 382)

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7, 1-26.

Fan, X., & Wang, L. (1995, April). How comparable are the jackknife and bootstrap results: An investigation for a case of canonical correlation analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 387 509)

Fish, L. J. (1988). Why multivariate methods are usually vital. Measurement and Evaluation in Counseling and Development, 21, 130-137.

Hinkle, D. E., Wiersma, W. & Jurs, S. G. (1994). Applied statistics for the behavioral sciences (3rd ed.). Boston: Houghton Mifflin.

King, J. E. (1997, January). Methods of assessing replicability in canonical correlation analysis. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Knapp, T. R. (1978). Canonical correlation analysis : A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Quenouille, M. H. (1949). Approximate tests of correlations in time-series. Journal of the Royal Statistical Society, 11, 68-84.

Stevens, J. (1996). Applied multivariate statistics for the behavioral sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Thompson, B. (1994a, February). Why multivariate methods are usually vital in research: Some basic concepts. Paper presented as a featured speaker at the biennial meeting of the Southwestern Society for Research in Human Development, Austin, TX. (ERIC Document Reproduction Service No. ED 367 678)

Thompson, B. (1994b, April). Common methodology mistakes in dissertations, revistied. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 367 678)

Thompson, B. (1995). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. Educational and Psychological Measurement, 55, 84-94.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25, 26-30.

Thompson, B. (in press). Canonical correlation analysis: Basic concepts and some recommended interpretaion practices. In L. Grimm & P. Yarnold (Eds.), Reading and understanding multivariate statistics (Vol. 2). Washington, DC: American Psychological Association.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. Annals of Mathematical Statistics, 29, 614.